

Multiword expressions in Russian thesauri RuThes and RuWordnet

Loukachevitch N., Lashevich G.

Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

© 2016 FRUCT. We present the types of multiword expressions included into the thesaurus of Russian language RuThes. Many of these expressions may look like compositional expressions but have specific relations that can be useful in applications. The relation system of the RuThes thesaurus allows natural description of relations between an expression and its components if necessary. Transforming the RuThes knowledge into the Princeton WordNet structure for creating Russian wordnet (RuWordNet), we transfer also all the described expressions into the new resource and propose to automatically introduce additional relations for their better representation.

References

- [1] E. Agirre, A. Izaskun Aldezabal and E. Pociello, "Lexicalization and multiword expressions in the Basque WordNet", in Proc. of Third International WordNet Conference, 2006.
- [2] L. Bentivogli and E. Pianta, "Extending Wordnet with Syntagmatic Information", in Proc. of Second Global WordNet Conference, 2004, pp. 47-53.
- [3] P. Kedzia, M., Piasecki, M. Maziarz and M. Marcirczuk, "Recognising compositionality of multi-word expressions in the wordnet oriented perspective", in Proc. of Mexican International Conference on Artificial Intelligence, 2013, pp. 240-251.
- [4] M. Maziarz, S. Szpakowicz and M. Piasecki, "A procedural definition of multi-word lexical units", in Proc. of Recent Advances in NLP Conference RANLP-2015, 2015, pp. 427-435.
- [5] C. Bontea, M. Green, J. Preciado and M. Palmer, "An approach to take multi-word expressions", in Proc. of the 10th Workshop on Multiword Expressions, 2014, pp. 94-98.
- [6] B. Dobrov and N. Loukachevitch, "Multiple Evidence for Term Extraction in Broad Domains", in Proc. of Recent Advances in NLP Conference RANLP-2011, 2011, pp. 710-715.
- [7] N. Astrakhantsev, D. Fedorenko and D. Turdakov, "Methods for automatic term recognition in domain-specific text collections: A survey", Programming and Computer Software, 2015, pp. 336-349.
- [8] A. Gelbukh and O. Kolesnikova, "Multiword Expressions in NLP: General Survey and a Special Case of Verb-Noun Constructions", in Computational Linguistics: Concepts, Methodologies, Tools, and Applications, 2014, pp. 178-198.
- [9] S. Gries, "50-something years of work on collocations: what is or should be next", International Journal of Corpus Linguistics vol. 18, 1, 2013, pp. 137-166.
- [10] E. Tutubalina, P. Braslavski, "Multiple Features for Multiword Extraction: Learning to Rank Approach" in Proc. of Conference on Computational linguistics and Intellectual technologies Dialog-2016, 2016, pp. 782-793.
- [11] P. Pecina, "Lexical association measures and collocation extraction", Language resources and evaluation vol. 44, 1-2, 2010, pp. 137-158.
- [12] M. Piasecki, M. Wendelberger, M. Maziarz, "Extraction of the Multiword Lexical Units in the Perspective of the Wordnet Expansion", in Proc. of Recent Advances in NLP Conference RANLP-2015, 2015, pp. 512-522.

- [13] S. Reddy, D. McCarthy and S. Manandhar. 2011. "An empirical study on compositionality in compound nouns", in Proc. of The 5th International Joint Conference on Natural Language Processing IJCNLP-2011, 2011.
- [14] S. Cordeiro, C. Ramisch, M. Idiart, A. Villavicencio, "Predicting the Compositionality of Nominal Compounds: Giving Word Embeddings a Hard Time", in Proc. of the 54th ACL-2016, Volume 1: Long Papers, 2016, pp. 1986-1997.
- [15] C. Ramisch, S. Cordeiro, L. Zilio, M. Idiart, A. Villavicencio, and R. Wilkens, "How naked is the naked truth a multilingual lexicon of nominal compound compositionality", in Proc. of the 55th ACL-2016, Volume 2: Short Papers, 2016.
- [16] Z39. 19. Guidelines for the Construction, Format and Management of Monolingual Thesauri. NISO, 2005.
- [17] G. Lebani and E. Pianta, " Encoding Commonsense Lexical Knowledge into WordNet", in Proc. of Global WordNet Conference, GWC-2012, 2012.
- [18] V. Vincze and A. Almasi, "Non-lexicalized concepts in wordnets: A case study of English and Hungarian", in Proc. of the 7th Global WordNet Conference GWC-2014, 2014, pp. 118-126.
- [19] Ruthes Thesaurus, Web: <http://www.labinform.ru/pub/ruthes/>
- [20] RuWordNet Thesaurus, Web: <http://ruwordnet.ru/>, <http://www.labinform.ru/pub/ruwordnet/index.htm>
- [21] N. Loukachevitch and B. Dobrov, "RuThes Linguistic Ontology vs. Russian Wordnets", in Proc. of Global WordNet Conference, GWC-2014, 2014, pp. 154-162.
- [22] N. Loukachevitch, B. Dobrov, " The Sociopolitical Thesaurus as a resource for automatic document processing in Russian", Terminology, 21 (2), 2015, 237-262.
- [23] N. Loukachevitch, G. Lashevich, A. Gerasimova, V. Ivanov and B. Dobrov. "Creating Russian WordNet by Conversion", in Proc. of Conference on Computational linguistics and Intellectual technologies Dialog-2016, 2016, pp. 405-415
- [24] N. Calzolari, Ch. Fillmore, R. Grishman, N. Ide, A. Lenci, C. MacLeod and A. Zampolli, " Towards Best Practice for Multiword Expressions in Computational Lexicons", in Proc. of LREC-2002, 2002, pp. 1934-1940.
- [25] I. Sag I, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger, " Multiword expressions: A Pain in the Neck for NLP", In Proc. of CICUNG 2002, 2002.
- [26] G. Miller, " Nouns in WordNet", In: Fellbaum, C (ed) WordNet-An Electronic Lexical Database, The MTT Press, 1998, pp. 23-47.
- [27] N. Guarino and C. Welty, "Evaluating ontological decisions with OntoClean", Communications of the ACM, V. 45 (2), 2002, pp. 61-65.
- [28] N. Loukachevitch and A. Alekseev, "Summarizing News Clusters on the Basis of Thematic Chains", in Proc. of Language Resources and Evaluation Conference, LREC-2014, 2014, pp. 1600-1607.
- [29] N. Loukachevitch and I. Chetviorkin, "Determining the most frequent senses using Russian linguistic ontology RuThes", in Proc. of the Workshop on Semantic resources and Semantic Annotation for Natural Language Processing and the Digital Humanities at NODAIADA 2015, 2015, pp. 21-27.